Exploring How Weak Supervision Can Assist the Annotation of Computer Vision Datasets

Andrea Abela
Department of Artificial Intelligence
University of Malta
Msida, Malta
andrea.abela.20@um.edu.mt

Dylan Seychell
Department of Artificial Intelligence
University of Malta
Msida, Malta
dylan.seychell@um.edu.mt

Mark Bugeja
Department of Artificial Intelligence
University of Malta
Msida, Malta
mark.bugeja@um.edu.mt

2022 IEEE 21st Mediterranean Electrotechnical Conference (MELECON) | 978-1-6654-4280-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/MELECON53508.2022.9843111

***Abstract*—Current artificial intelligence (AI) workflows depend on researchers performing laborious annotation work. In the case of computer vision, crowdsourcing is a popular alternative to alleviate this effort. The general public can provide even more trustworthy annotations with the help of existing frameworks. This paper proposes an image dataset annotator helper that uses weak supervision and explains how class activation maps (CAMs) are integrated with deep image classifiers to produce weakly supervised localisers that could further improve human image annotation performance. Comparing these models with primary crowdsourcing data revealed that the models can annotate better than humans by 9.7% when measuring the localisation error while taking into account both false positives (FPs) and false negatives (FNs). Moreover, the models can also save up to 36% of the time required to perform manual image annotation. This confirms that there is potential within CAM-empowered models to further improve the image annotation experience.**

***Index Terms*—artificial intelligence, computer vision, convolutional neural networks, weak supervision, object localisation**

I. Introduction

Today, most modern artificial intelligence (AI) techniques predominantly leverage information from data to perform simple and complex tasks. This reliance on data affects the model's performance, with specific tasks requiring more data than others [1]. These datasets are frequently manually annotated, especially when dealing with real-world problems, where simulated data seldom yields the granularity necessary to perform well. The data gathering and annotation process is time-consuming, requires many human resources, and is prone to human error. Work has been developed that allows for the rapid creation of datasets and annotations built through crowdsourcing technology, notably [2], [3]. Such techniques, although practical, raise questions on reliability, especially for tasks where the data annotations from different sources do not reach the level of statistical confidence required to ensure that the dataset is accurate. A problem such as this is critical in machine learning (ML) and computer vision (CV). Unreliable data can yield models that perform well on the testing set but poorly on the validation set or, when plugged in, in the wild [4]. Digital systems that dealt with image annotation, such as Snuba, were used to improve such annotations while also keeping the act of crowdsourcing relevant [5]. However, such systems needed significant effort. For instance, Snorkel and Snuba required participants to program the conditions for the model used [5], [6]. This paper introduces a method that builds upon this work by adding a dataset annotator helper that uses weakly supervised learning. Weakly supervised models empowered by class activation maps (CAMs) were compared with primary data obtained from tailor-made solutions that imitate crowdsourcing. This experiment assessed the reliability of machine-generated annotations when compared to human annotations. In this experiment, a web form was implemented to help measure human reliability in determining which annotations were manufactured and report the effectiveness of using CAMs as an aid in dataset annotation.

II. Literature Review

A. Image classification

ML models known as neural networks (NNs) consist of virtual neurons that process the input data's extracted features, similar to the inner workings of a brain. These NN-based architectures proved to be competent as they were used in critical scenarios like breast cancer classification [7]. However, ML methods like the NN require much maintenance due to their dependency on manual feature extraction. With the emergence of deep learning (DL) and convolutional neural networks (CNNs), these data sanitisation prerequisites were made redundant [8]. Early architectures like AlexNet brought on further ease of use to DL by empowering training sessions with the graphical processing unit (GPU) [9]. In addition to AlexNet, more architectures such as VGG [10], MobileNet [11], and EfficientNet [12] advanced the state of the art by continuously adding new features or optimising their performance. CNNs have also been used in many critical applications like breast tumour segmentation [13] and garbage classification [14].

B. Object detection

ML object detection methods included techniques like histogram of oriented gradients (HOG) [15] and Haar cascades [16]. Similar to image classification, these methods also required manual feature extraction. With the inclusion of DL, however, more accessible and powerful architectures like feature pyramid networks emerged [17]. Such models are known for their effective localisations as they use a two-stage approach [17]. In contrast, one-stage approaches, like

SSD, are renowned for their efficiency as they can reach inference speeds of up to 46 frames per second (FPS) [18]. The utilisation of deep image classifiers as a backbone within object detection architectures also resulted in more efficient models [11], [18]. Object detection has been used in important use cases like litter detection [19] and vehicle detection at night [20]. Bounding box annotations on large samples of data, however, still impede the maintainability of deep object detectors [8].

C. Dataset annotation

Despite DL's advantages, dataset annotation is still a relatively laborious task. The dataset known as Conceptual Captions contains reported that only 3% of the automatically acquired images fit their use case [21]. Large crowdsourced datasets like ImageNet have many overly-specific and identical labels which could have confused its participants [22]. Furthermore, even professionally annotated datasets like Google's Open Images exhibit inaccuracies as the dataset known as More Inclusive Annotations for People (MIAP) added 100,000 missing "person" annotations [23]. This study notes how common inconsistent dataset annotation methods are, even when they serve similar purposes [24]. This same study also mentions the importance of third-party annotations, similar to the concept of crowdsourcing [24].

D. Supervised techniques

1) Fully supervised learning: Within the domain of object detection, a fully supervised dataset provides both the labels and bounding boxes. This is the most straightforward instance of fully supervised learning as previously reviewed models utilise this technique [10]–[12].

2) Unsupervised learning: The absence of annotations within structured data can still be used to discover trends and relationships between classes [25]. This technique, known as unsupervised learning, is only applicable to clustering or association scenarios. In CV, it is applied to image processing techniques like denoising [26] and compression [27].

3) Self-supervised learning: Self-supervised learning leverages unlabelled data in an unsupervised dataset to use within classification scenarios. While it has been used within CV, it is currently computationally expensive and inaccurate [28], [29].

4) Weakly supervised learning: Weakly supervised learning is a technique that consists of many paradigms [30]. This group of techniques attempts to maximise model performance through the usage of vague labels. The vagueness of these labels varies depending on the paradigm used [30].

a) Heuristic frameworks: Heuristic rules are solutions to computationally complex problems that sacrifice either precision, optimisation or completeness. This technique can be used to automatically label missing training data via a subject expert. Snuba and Snorkel are frameworks that use this methodology for weakly supervised learning [5], [6]. However, they require much effort as the subject expert has to program the conditions themselves.

b) Multiple instance learning: Multiple instance learning (MIL) is a weakly supervised technique that can been applied to CV dataset annotation [31], [32]. MIL organises unlabelled samples into different cropped image sections called bags, which contain positive and negative instances of the label. As a result, this method is mainly used in binary classification.

c) Semi-supervised learning: Semi-supervised learning is utilised when not all the training samples are annotated. This technique learns the partially labelled dataset's features to infer the annotations of the unlabelled data [30]. Semi-supervised learning has been used within CV but has been reported to be unstable [33].

d) Class activation maps: CAMs are an AI explainability method that can be used for weakly supervised learning. There exist many variations, each with differing methodologies and results. Figure 1 presents some common variations. Grad-CAM and Grad-CAM++ use the throughput's gradients to track the most influential features [35], [36]. In contrast, Score-CAM and Faster Score-CAM track the individual weights of the model through forward passes [34]. Additionally, weakly supervised models empowered by CAMs operate similarly to object detectors such as SSD, which use image classifiers as a base network [18].

III. Methodology

Due to their potential in reducing both time and effort, CAMs were used to develop a basis for dataset annotator helpers. This implementation was further guided by a pipeline that is organised into individual modules, namely "Crowdsourcing emulation", "Annotations comparison" and "Human reliability evaluation". This pipeline is illustrated in Figure 2. Each module serves to answer each objective, respectively.

A. Crowdsourcing emulation

The first module concerned the development of a crowdsourcing solution in the form of an interactive web application. This web application provided all functions of an image annotator, focusing on user performance tracking. The dataset used throughout this study was a simplified version of ImageNet [37]. ImageNet was superclassed through the Robustness[1]

[1]https://robustness.readthedocs.io/en/latest/

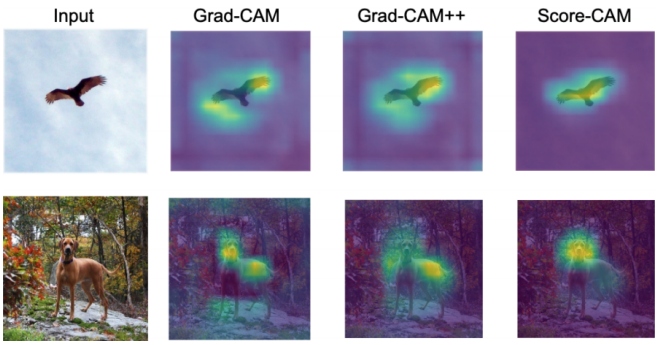


Fig. 1: Varied discrimination and localisation abilities of different CAM techniques on images [34]

package via the "big_12" configuration. Subsequently, this process yielded a more legible dataset with a smaller number of labels. The survey was split into two parts: classification and localisation. Respective to their task, classification allows users to label the images, while localisation allows users to both label images and add bounding boxes. Since the dedicated survey subset was still relatively large, a load balancing system similar to this study's was implemented to limit the bias of the acquired data [24]. All data was then saved in a database for further processing.

B. Annotations comparison

The training subset for the simplified ImageNet dataset was used to train the core AI models. The CNN architectures were chosen based on whether they presented a localisation score or have been commonly used as a base network within object detectors. Additionally, they must have been pre-trained on ImageNet to maximise their performance. These conditions pruned the eligible image classifiers to the following models:

- **VGG16**: Presented a localisation score and has seen use in object detection [10], [18].
- **MobileNetV2** (MNV2): Designed for mobile devices and is commonly used in conjunction with SSD [11].
- **EfficientNetB0** (ENB0): Efficient but accurate baseline architecture that used CAMs for its evaluation. It is used in both classification and detection [12], [38].

To complete the weakly supervised workflow, some CAM methods were also selected to complement the chosen models. These include:

- **Grad-CAM** (GC): Utilises model gradients to improve the generalisation of the result while also removing the need for global average pooling (GAP) layers [35].
- **Grad-CAM++** (GC++): Claims to improve on Grad-CAM by using techniques like pixel-wise weighting [36].
- **Score-CAM** (SC): Claims to remove the instability of gradients in favour of a score-based approach [34].
- **Faster Score-CAM** (FSC): Improves on Score-CAM efficiency by limiting the extraction of effective maps[2].

Each model architecture was branched to provide two results. The first branch outputted the predicted label while the second branch outputted the input's features. The input's features were fed to the relevant CAM technique to generate a heatmap based on the model's predicted label. This heatmap was then converted to bounding box data by using Otsu's threshold [39]. The TensorFlow framework was used to deploy this implementation's models. Additionally, KerasTuner was responsible for determining the ideal configuration for the training sessions [40]. KerasTuner is a tool that uses an oracle system to aid researchers in finding the optimal hyperparameters for AI models. The HyperBand oracle was used on a small subset of the training set. The optimisers used were based on the architectures' original papers.

[2]https://github.com/tabayashi0117/Score-CAM

Configuration	Label$_{opt}$	Label$_{full}$	Localise$_{opt}$	Localise$_{full}$
E	10s	17s	17s	23s
$(E-U)$	**9s**	**15s**	**14s**	**22s**
U	18s	31s	23s	30s

TABLE I: Human annotation times separated by survey parts with times presented as n_m where n is the annotation type and m is whether it is the full or optimised time. E represents all entries and U represents all entries with an unknown label. The shortest times are marked in bold.

C. Human reliability evaluation

Evaluating both the man-made and machine-generated annotations on the ground truth gave rise to a contradiction. Since the original ground truth annotations were annotated by humans, then the results are biased towards humans. Therefore, a secondary online survey was developed to further support the comparisons. This web form tested the participants on their ability to recognise man-made and machine-generated annotations. The results of this form verified the machine-generated annotations' integrity without directly using the ground truth. In addition to the test, questions related to the participant's demographic and sentiment were included to supplement the results.

IV. Results

A. Crowdsourcing intervals

1) Labelling vs localising: The data recorded in the crowdsourcing experiment tracked the performance of many users through user interface (UI) event timestamps spanning 5606 entries. Table I shows the optimised and full average entry times by entry type configuration. The optimised times were calculated by negating the time taken to perform unnecessary actions from the full entry time. Set E represents all entries while set U represents all the entries lacking a label. It can be concluded from Table I that labelling takes less time to perform than localising. Consequently, this alludes to the possibility that a weakly supervised workflow could be more efficient than a fully supervised workflow.

2) Human inefficiency: By using the same timestamps, it was also discovered that participants spent at least 30% of the time not performing annotations. If one were to exclude computational delays, a digital system would not suffer from such inactivity rates.

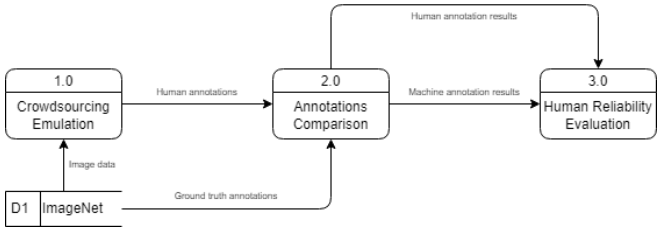


Fig. 2: Research pipeline divided into modules

Configuration	LE	LE_{fp}	LE_{fn}	LE_{fp+fn}
E	43.9%	53.5%	50.5%	58.9%
$E - U$	37.5%	48.2%	44.8%	54.2%
$\approx E$	34.2%	44.2%%	42.5%	49.2%
$\approx (E - U)$	**31.7%**	**42.5%**	**40%**	**48.3%**

TABLE II: Human annotation results by localisation error (LE) variations which take into account either false positives (LE_{fp}), false negatives (LE_{fn}) or both (LE_{fp+fn}). E represents all localisation entries while U represents all localisation entries with an unknown label. $\approx n$ represents the approximated entries by image where n is any set of entries. The best results are marked in bold.

3) Annotation preferences: Further analysing the intervals revealed that 34% of the localisation part's entries were created via the label select element first. This is relatively similar to the inner workings of CAMs, where the localisations are generated through the predicted label. This also suggests that the label is a determining factor for annotations, both for humans and models.

B. Human vs machine metrics

1) Hyperparameter tuning and training: Preparing the models for training consisted of supplying KerasTuner with different learning rate and momentum configurations. Let set $L = \{1 \times 10^{-2}, 1 \times 10^{-3}, 1 \times 10^{-4}, 1 \times 10^{-5}\}$ represent all possible learning rates and set $M = \{0, 0.9\}$ represent all possible momentum values. After completion, KerasTuner yielded the following recommendations where l is the best learning rate value from set L while m is the best momentum value from set M:

- **VGG16**: $l = 1 \times 10^{-3}$, $m = 0.9$
- **MobileNetV2**: $l = 1 \times 10^{-4}$, $m = 0.9$
- **EfficientNetB0**: $l = 1 \times 10^{-3}$, $m = 0.9$

The deep image classifiers were trained with these configurations. Following the training sessions, the models were integrated with each CAM technique, to produce a combination of results.

2) Metrics: Since ImageNet was used as a basis for this study, the localisation error (LE) metric from the ILSRVC was used to present the both the weakly supervised model and the survey entry results [37]. However, the LE metric exhibited a number of oversights. Since the result was determined by finding at least one match, then the LE metric commonly reported inaccurate results. Variations of the LE metric were developed to glean better insight into the results. Tables II and III present these new metrics alongside their original counterpart. Table II shows that the survey participants performed averagely in determining the ground truth. With the lowest error of 31.7% only having been achieved by rounding all entry results by their image, removing the unknown entries, and using the lenient LE metric. In contrast, the LE_{fp+fn} metric reveals how biased the original LE metric is, with a 15.8% average difference between LE and LE_{fp+fn}. The same phenomenon can be observed in Table III, where the model-CAM combinations were subjected to the same process. Many of the models performed worse than the survey participants, with most of the LE_{fp+fn} metrics being over 60%. However, MobileNetV2 integrated with Faster Score-CAM outperformed all the other model-CAM combinations by a significant amount. Additionally, it also exceeded over the performance of the unprocessed survey results by 9.7% LE_{fp+fn} and almost reach the fully processed survey results within 0.9% LE_{fp+fn}. Moreover, if LE_{fp+fn} was not utilised, VGG16 would have been nominated as the optimal model. This further justifies the usage of the LE variations.

3) Confidence threshold: To further equalise the results, the model-CAM combinations also had access to a "none of the above" label. If the result was accompanied by a confidence score $c < 0.5$, then the output would have been modified to a "none of the above" label instead. Despite this change, none of the AI models returned this label, which suggests that the AI models are more "confident" than humans within that threshold.

4) Pseudo-sessions: Furthermore, by using the training and inference times of the model-CAM combinations, and the total amount of training images within the training set of the simplified ImageNet, the annotation times could be compared with the survey participants' time intervals. It was discovered that the AI models could save between 17 and 36% of the time required for manual annotations. This further implies their utility within this context.

C. Annotation choices assessment

1) Demographics: After accumulating 119 responses from the web form, the data was aggregated. Most participants were 26 - 35 years old, while there was an even distribution in the other ranges. Additionally, about two-thirds of the participants were men. Categorising the participants' annotation confidence showed that users were very confident in their abilities. In fact, the arithmetic mean of the annotation confidence score was 84%.

2) Annotation choice results: In the test, the participants obtained 36.7% average accuracy. Since the form contained

Model	CAM	LE	LE_{fp}	LE_{fn}	LE_{fp+fn}
ENB0 [12]	GC [35]	68.3%	71.7%	71.7%	74.2%
ENB0 [12]	GC++ [36]	64.2%	69.2%	67.5%	71.7%
ENB0 [12]	SC [34]	65.8%	69.2%	68.3%	71.7%
ENB0 [12]	FSC [34]	65.8%	66.7%	70.8%	70.8%
MNV2 [11]	GC [35]	60%	65%	63.3%	67.5%
MNV2 [11]	GC++ [36]	55%	58.3%	58.3%	60.8%
MNV2 [11]	SC [34]	54.2%	57.5%	58.3%	60.8%
MNV2 [11]	FSC [34]	46.7%	**46.7%**	49.2%	**49.2%**
VGG16 [10]	GC [35]	74.2%	82.5%	78.3%	85%
VGG16 [10]	GC++ [36]	**40.8%**	55.8%	**48.3%**	60%
VGG16 [10]	SC [34]	44.2%	55%	53.3%	60.8%
VGG16 [10]	FSC [34]	56.7%	62.5%	61.7%	66.7%

TABLE III: Model-CAM results by localisation error (LE) variations which take into account either false positives (LE_{fp}), false negatives (LE_{fn}) or both (LE_{fp+fn}). The best results are marked in bold.

two correct answers per question and the participants were required to choose two answers per question, this reveals that the participants were more likely to choose a machine annotation as a human annotation. Additionally, only 12.5% of the participants managed to answer over half of the annotation questions correctly. Some trends were also observed from the aggregated data. Namely, participants were more likely to choose the annotations that were perceived as correct. This reveals that the typical participant equates a human annotation as a good annotation. Some participants also determined their answer by deliberately choosing the incorrect answer. More specifically, Figure 3a presents an image where the ground truth shows that the correct answer is "furniture". However, the survey participant localised the sock as "clothing", whereby the form participants realised that such performance was distinguishable as human. Additionally, further alternative behaviour was observed when participants dismissed the obviously incorrect annotations as machine annotations. This further strengthens the argument that the typical participant assumed that a good annotation can only be made by a human.

3) Sentiment: Finally, the sentiment-oriented questions were aggregated. Participants reported an average difficulty rating of 65% in determining the annotation choices of the previous section. If one were to consider this question as the inverse of the annotation confidence question, the participants' confidence seemed to have decreased. This alludes to the possibility that the participants had a change of opinion after completing the form. Despite this, this difficulty rating is relatively low when considering the actual accuracy of the participants' annotation choices, further showing how humans are inconsistent within image annotation work. By inference, this suggests that the machine annotations are truly comparable to human annotations. In fact, the final question dealt with the form participants' opinion with dataset annotator helpers. Aggregating these responses revealed that 85% of the participants showed positive sentiment towards such software. Conversely, 15% of the participants responded negatively to such a system. However, two-thirds of this segment of responses remarked that it is possible that they would change their opinion if accuracy is guaranteed with this system.

V. Conclusion

This paper investigated the effectiveness of CAM-empowered weakly supervised models as dataset annotator helpers. This was achieved by analysing the inferences of the models with crowdsourcing results achieved through an interactive web application. These results were further supported by an auxiliary form that showed the unreliability of humans within image annotation. The weakly supervised models outperformed the unprocessed human annotations (E) in terms of LE_{fp+fn} by 9.7% while also saving between 17 and 36% of the time, showing that CAMs can be effective dataset annotator helpers.

References

[1] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting unreasonable effectiveness of data in deep learning era," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 843–852, 10 2017.

[2] B. M. Good, S. Loguercio, O. L. Griffith, M. Nanis, C. Wu, and A. I. Su, "The cure: Design and evaluation of a crowdsourcing game for gene selection for breast cancer survival prediction," *JMIR Serious Games*, vol. 2, p. e7, 07 2014.

[3] J. C. Chang, S. Amershi, and E. Kamar, "Revolt: Collaborative crowdsourcing for labeling machine learning datasets," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 2334–2346, Association for Computing Machinery, 05 2017.

[4] N. Sambasivan, S. Kapania, H. Highfill, D. Akrong, P. Paritosh, and L. Aroyo, ""everyone wants to do the model work, not the data work": Data cascades in high-stakes ai," in *CHI '21: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–15, Association for Computing Machinery, 05 2021.

[5] P. Varma and C. Ré, "Snuba: Automating weak supervision to label training data," *Proceedings of the VLDB Endowment*, vol. 12, pp. 223–236, 11 2018.

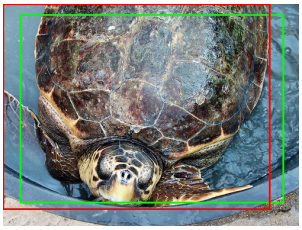
[6] A. Ratner, S. H. Bach, H. Ehrenberg, J. Fries, S. Wu, and C. Ré, "Snorkel: rapid training data creation with weak supervision," *The VLDB Journal*, vol. 29, pp. 709–730, 05 2020.

[7] F. F. Ting and K. S. Sim, "Self-regulated multilayer perceptron neural network for breast cancer classification," in *2017 International Conference on Robotics, Automation and Sciences (ICORAS)*, pp. 1–5, 11 2017.

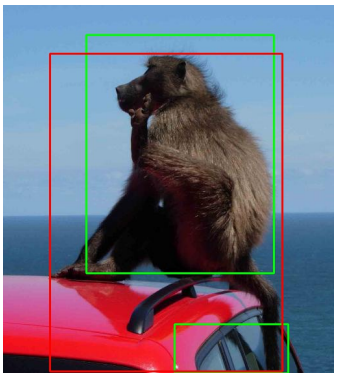
[8] N. Buduma and N. Locascio, *Fundamentals of Deep Learning : Designing Next-generation Artificial Intelligence Algorithms*. O'Reilly Media, Inc., 2017.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Infor-*

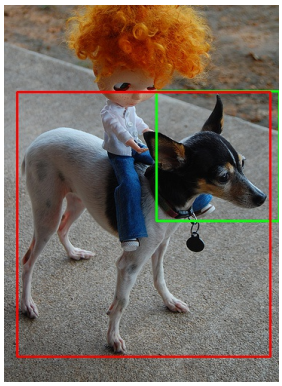
(a) Survey participant localised "clothing" when truth label is "furniture" referring to the entire bed



(b) Survey participant localised "none of the above" when truth label is "reptile"



(c) VGG16 with Grad-CAM++ localised "primate" when bounding box truth outlines entire monkey in only one



(d) VGG16 with Grad-CAM localised "dog" when bounding box truth outlines entire dog rather than only its face

Fig. 3: Various incorrect annotation samples from both survey participants and models where ground truth is denoted in red and green denotes the prediction

mation Processing Systems, vol. 25, 01 2012.
[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 04 2015.
[11] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4510–4520, 06 2018.
[12] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the 36th International Conference on Machine Learning*, vol. 97, pp. 6105–6114, PMLR, 06 2019.
[13] N. Micallef, D. Seychell, and C. J. Bajada, "Exploring the u-net++ model for automatic brain tumor segmentation," *IEEE Access*, vol. 9, pp. 125523–125539, 2021.
[14] A. Abela and T. Gatt, "Using class activation maps on deep neural networks to localise waste classifications," in *2021 IEEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMI)*, pp. 000143–000148, 01 2021.
[15] T. Watanabe, S. Ito, and K. Yokoi, "Co-occurrence histograms of oriented gradients for human detection," *IPSJ Transactions on Computer Vision and Applications*, vol. 2, pp. 39–47, 2010.
[16] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, vol. 1, p. 1, 02 2001.
[17] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 07 2017.
[18] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Computer Vision – ECCV 2016*, pp. 21–37, Springer International Publishing, 2016.
[19] M. Schembri and D. Seychell, "Small object detection in highly variable backgrounds," in *2019 11th International Symposium on Image and Signal Processing and Analysis (ISPA)*, pp. 32–37, 09 2019.
[20] S. Galea, D. Seychell, and M. Bugeja, "A survey of intelligent transportation systems based modern object detectors under night-time conditions," in *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, pp. 265–270, 12 2020.
[21] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2556–2565, Association for Computational Linguistics, 07 2018.
[22] L. Beyer, O. J. Hénaff, A. Kolesnikov, X. Zhai, and A. v. d. Oord, "Are we done with imagenet?," 06 2020.
[23] C. Schumann, S. Ricco, U. Prabhu, V. Ferrari, and C. Pantofaru, "A step toward more inclusive people annotations for fairness," in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 916–925, 07 2021.
[24] D. Seychell, C. J. Debono, M. Bugeja, J. Borg, and M. Sacco, "Cots: A multipurpose rgb-d dataset for saliency and image manipulation applications," *IEEE Access*, vol. 9, pp. 21481–21497, 2021.
[25] G. Caruso and S. A. Gattone, "Waste management analysis in developing countries through unsupervised classification of mixed data," *Social Sciences*, vol. 8, p. 186, 06 2019.
[26] S. Kuanar, V. Athitsos, D. Mahapatra, K. Rao, Z. Akhtar, and D. Dasgupta, "Low dose abdominal ct image reconstruction: An unsupervised learning based approach," in *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 1351–1355, 09 2019.
[27] J. Song, T. He, L. Gao, X. Xu, A. Hanjalic, and H. T. Shen, "Unified binary generative adversarial network for image retrieval and compression," *Int J Comput Vision*, vol. 128, pp. 2243–2264, 09 2020.
[28] J. Jiao, R. Droste, L. Drukker, A. T. Papageorghiou, and J. A. Noble, "Self-supervised representation learning for ultrasound video," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pp. 1847–1850, 04 2020.
[29] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *Computer Vision – ECCV 2016*, pp. 69–84, Springer International Publishing, 2016.
[30] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *Natl. Sci. Rev.*, vol. 5, pp. 44–53, 01 2018.
[31] G. Xu, Z. Song, Z. Sun, C. Ku, Z. Yang, C. Liu, S. Wang, J. Ma, and W. Xu, "Camel: A weakly supervised learning framework for histopathology image segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 10 2019.
[32] J. Correia, I. Trancoso, and B. Raj, "Automatic in-the-wild dataset annotation with deep generalized multiple instance learning," in *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 3542–3550, European Language Resources Association, 05 2020.
[33] J. E. van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Mach Learn*, vol. 109, pp. 373–440, 02 2020.
[34] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, and X. Hu, "Score-cam: Score-weighted visual explanations for convolutional neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 24–25, 06 2020.
[35] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, 10 2017.
[36] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 839–847, 03 2018.
[37] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: Constructing a large-scale image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.
[38] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10781–10790, 06 2020.
[39] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern.*, vol. 9, pp. 62–66, 01 1979.
[40] T. O'Malley, E. Bursztein, J. Long, F. Chollet, H. Jin, L. Invernizzi, and K. Team, "Keras tuner," 2019.